University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 8: k-Nearest Neighbor Classification, Introduction to R

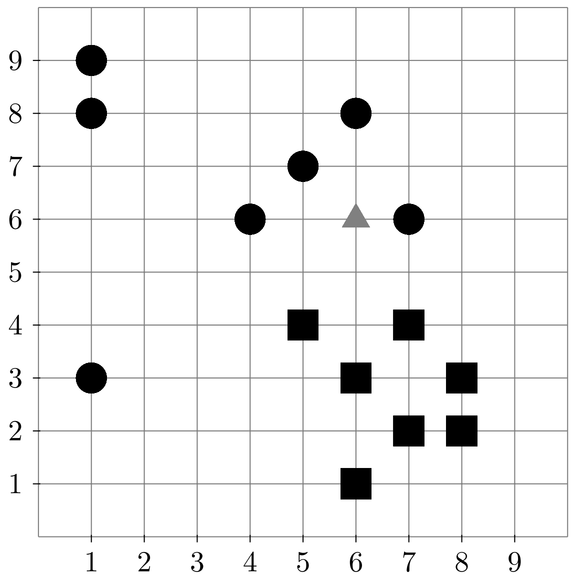
Exercise 8-1 Nearest neighbor classification

The 2D feature vectors in the figure below belong to two different classes (circles and rectangles). Classify the object at $(6, 6)$ — in the image represented using a triangle — using k nearest neighbor classification. Use Manhattan distance (L_1 norm) as distance function, and use the non-weighted class counts in the k-nearest-neighbor set, i.e. the object is assigned to the majority class within the k nearest neighbors. Perform kNN classification for the following values of k and compare the results with your own "intuitive" result.

(a) $k = 4$

(b) $k = 7$

(c) $k = 10$



Exercise 8-2 Nearest Neighbor classification

Give a set of points, consisting of at least four points in 2 dimensions, such that the Nearest Neighbor classification ($k = 1$) only gives incorrect classification results. Use Euclidean distance as distance function.

Exercise 8-3 Get started with R

(a) Download and install R-Studio, so you are ready for working in R on your laptop when you arrive in class: `https://www.rstudio.com/products/rstudio/download/`.

You can try the following exercise suggestions yourself (and explore more of R as much as you want). However, the following exercises will be performed step-by-step interactively in the exercise class.

(b) Start a new R script containing your solutions. Save the script for later reference.

You can use #### Section With Name #### To define a section within your script with name "Section With Name" in your R code, that you can easily navigate to.

(c) Some important commands for learning R, are `help()`, `class()`, and `mode()`.

You can use these commands on variables, functions, objects, and datasets to obtain information on them.

Exercise 8-4 Vectors in R

(a) Create a vector of length 5 containing both positive and negative numbers, using the concatenate (`c()`) command.

(b) Find the `mean()`, `max()`, `min()` of the vector. Then compute the mean of the absolute values.

(c) Taking a subset of the vector can be done using the following notation:

`vector[1:2]` will take the first two elements of the vector (*R* starts indexing with 1).

Insert 42 on the third position of the vector you created earlier.

(d) Create a new vector and build the sum of the two vectors.

(e) Create a random vector using the `rnorm()` function with no additional arguments.

- Calculate the mean — what do you observe?
- Take the last 5 elements of the vector using the indexing described above.

Exercise 8-5 Matrices in R

(a) Create a 2×2 matrix A by row binding vectors using the `rbind()` command.

(b) Nullify matrix A by adding another matrix that you define.

(c) Double all the values in the original matrix A by multiplication with another matrix that you define.

Exercise 8-6 Exploration of Datasets in R

(a) Use the help command to get information on the built-in dataset AirPassengers.

   (i) Plot the dataset using the `plot()` command. What do you see? Describe the resulting plot.

   (ii) Create a histogram using the built-in function `hist()`
   What do you observe?

   (iii) Check the `class()` and `mode()` of the dataset. Are these as expected? If you are not sure what the `mode` and `class` functions do use the `help()` function.

(b) *R* comes with many historical data sets. One of them is the Titanic dataset. Use the `help()` function to read about the data set. Then make a mosaic plot using the `mosaicplot()` command. What do you observe?

Exercise 8-7 Clustering and Classification on the Iris Data

(a) Load the Iris dataset in R, remove the class attribute. Cluster it using k-means with a reasonable choice of k.

(b) Use the clustering result to label the data.

(c) Create some artificial flower data, that could potentially be Iris flowers.

Think about how you will do this, and what you expect the resulting flowers to be.

(d) Try the $k$nn-classifier with different values for k, and use your generated labeled Iris dataset to classify the artificial query points.

(e) Try using the original labeled Iris dataset. Does this yield the same result?

(f) Explain your findings.